EXPLORATION OF CRASH PRONE TRAFFIC CONDITIONS ON FREEWAYS IN THE NETHERLANDS USING RANDOM FORESTS

Mohamed Abdel-Aty, Anurag Pande, Abhishek Das, Willem Jan Knibbe

ABSTRACT: Underground loop detectors are one of the most common traffic surveillance apparatus on freeways. Traffic data obtained from these detectors are used for various ITS (Intelligent Transportation Systems) applications such as travel time estimation and incident detection. In the recent past, however, researchers have been interested in proactive applications of these data. These proactive applications primarily involve real-time crash risk assessment based on analysis of traffic surveillance data observed prior to historical crashes. In this study, we have analyzed the crash data from five freeway sections in the Utrecht region of the Netherlands to identify the traffic conditions significantly associated with crash occurrences. Random Forest, a data mining methodology employing multiple classification trees, would be used to analyze the data and identify traffic parameters significantly associated with the binary variable representing crash vs. non-crash. It was found that the turbulence in traffic speeds is related to real-time crash likelihood. The results are promising in that they show the potential of transferring the proactive traffic management approach proposed by Pande and Abdel-Aty (2007) for Interstate-4 in Orlando, FL to freeways in the Netherlands.

INTRODUCTION

Underground loop detectors are used to continuously collect real-time traffic information from highways (primarily uninterrupted flow facilities such as freeways) around the World. Data from these detectors are used for various ITS (Intelligent Transportation Systems) applications such as travel time estimation and incident detection. Recently, however, the focus has been shifting towards proactive applications of these data including the development of real-time crash risk assessment models. These models are developed based on analysis of traffic surveillance data observed prior to historical crashes. For example, Golob and Recker (2004) and Pande and Abdel-Aty (2006 a) related crash patterns to traffic data from single loop detector stations in California and dual loop detector stations in Florida, respectively. These models can differentiate crash prone traffic conditions from non-crash (or 'normal') traffic conditions in real-time. They can potentially be used to issue warnings or implement traffic management strategies ahead in time for reducing measure(s) of crash risk and avoid an imminent crash.

Despite the extensive analysis presented in these studies a critical issue that remains to be addressed is transferability of such an approach especially to locations with varying traffic surveillance systems. The archived freeway traffic data may vary in terms of collected traffic parameters, level of aggregation, and/or spacing of loop detectors. In this study we would explore a sample of crash as well as non-crash data from five freeways in the Utrecht region of the Netherlands.

When compared to most freeways in the US the traffic surveillance/management apparatus on these freeways is more advanced. Existing application of variable speed limits through Dynamic Message Signs is one such advancement. The speed limits are implemented either automatically (primarily in response to congestion triggers) or manually by operators. It would be of interest to understand what impact, if any, this implementation has on real-time safety. On the Dutch freeways the density of loop detectors is higher, i.e., the detectors are more closely spaced than the comparable traffic surveillance systems in the US while the distance between them is more variable. The data collected from these detectors were available to us in the form of 1-minute aggregates as compared to 30-seconds data used in our previous studies.

In this study a relatively recent data exploration technique, namely, Random Forests, has been used to identify significant variables instead of the classification trees. Random Forests, which is a combination of multiple tree classifiers, tends to be more robust compared to classification trees. It works efficiently on large datasets and is being increasingly used for variable selection. The study is one of the first applications of Random Forests based variable selection in the field of transportation engineering. Real-time traffic parameters significantly associated with crashes on Dutch freeways have been contrasted with the results from our previous studies on Interstate-4 in Orlando, FL (USA). The comparison provides insight into issues associated with transferability of the approach to real-time crash risk assessment.

DATA COLLECTION AND PREPARATION

Study Area and Available Data

As mentioned earlier, data from loop detectors on five freeway sections in the Utrecht region of the Netherlands were used in this study. The freeways and mileposts (in km) corresponding to the extremities of the sections are as follows: A1 (28.82 km – 46.96 km), A12 (39.8 km – 91.6 km), A2 (36.9 km – 94.58 km), A27 (53.35 km – 100.35 km) and A28 (0.25 km – 32.375 km). Loop

detectors measuring the traffic parameters are present on both direction of travel even though they are not perfectly aligned at certain locations. Also, note that spacing between consecutive loop detectors in the same direction of travel is not consistently same but is always less than 0.8 km (0.5 mile). Hence, the arrangement is significantly different than the series of loop detectors present on Interstate–4 which was the freeway used in one of our previous studies [2]. As noted later in the paper this difference is critical in devising the approach to data analysis.

The following traffic parameters are available for every minute; speed, volume (normalized as hourly volume), a flag for traffic congestion along with the message displayed on the dynamic message signs (DMS). The possible DMS displays include the implemented variable speed limits or suggested traffic maneuvers (e.g., merge left, etc.). Note that the dataset does not contain information on lane-occupancy. These data are available for the month of September 2006 along with the incident reports for the same period and freeways. The incident reports included 288 crashes; loop data corresponding to which were extracted and used in this study.

Data Preparation

First, the location and time of occurrence for each of the 288 crashes were identified. Then for every crash, six loop detectors stations (three stations in the upstream direction and three in the downstream direction) were identified. Since some of the loop detectors were very closely spaced it was decided to have a minimum of 250 m distance between consecutive loop detectors stations. The threshold of 250 m was established because of the arrangement of loop detectors. A lower threshold would have meant that the data from two consecutive loop detectors would provide scant independent information. A higher threshold (say ½ mile = 0.8 km comparable to the average distance between consecutive loop detectors on Interstate-4 in Orlando, FL) would have meant that the six detectors from which data would be collected are spread around 5 to 6 km (2.5 to 3 km in each direction). That would in turn mean that there must be about 2.5 to 3 km instrumented section (section with loop detectors installed) of the freeway on both sides (upstream and downstream) of crash location. Hence, crashes that occurred within almost 2.5 to 3 km of either boundary on either of the five freeway segments would have required traffic data missing. A larger threshold such as 0.8 km would essentially mean leaving out more crashes closer to the two ends of the five freeway corridors. Since we are dealing with five different roadways rather than one long freeway section, this would have resulted in significant reduction in our already limited sample size.

The next step was to extract loop data corresponding to the crashes. The first upstream and downstream loop detectors stations relative to the crash were named 'US1' and 'DS1', respectively. The subsequent loop detector stations in either direction were named 'US2'/'DS2', and 'US3'/'DS3', respectively. Figure 1 shows the positions of the upstream and downstream loop detector stations relative to a crash and the minimum spacing between them. In Figure 1, Position 1 represents the set of two loop detectors that are nearest to crash location in the upstream and downstream direction. Position 2 and Position 3 include the sets of two subsequent detectors in upstream and downstream directions. The significance of defining these sets of detectors would be clear towards the end of this section. The loop data were then extracted in a following format. If a crash, for example, has occurred on September 4, 2006 (Monday) at 05:00 p.m. on the freeway A2, then the corresponding loop detectors stations of interest are 'US1', 'US2', 'US3' in the upstream direction and 'DS1', 'DS2', and 'DS3' in the downstream direction. This crash case will have a loop database consisting of 1-minute averages of speed and

volume, congestion flag along with the message displayed on the DMS for that minute for all the lanes at the six stations from 04:50 p.m. to 05:05 pm (15 minute window) on September 4, 2006. A variable 'Y' was created with value as 1 for all the crashes. It would be used as the binary target variable for the Random Forests with its value as 0 for the non-crash cases.

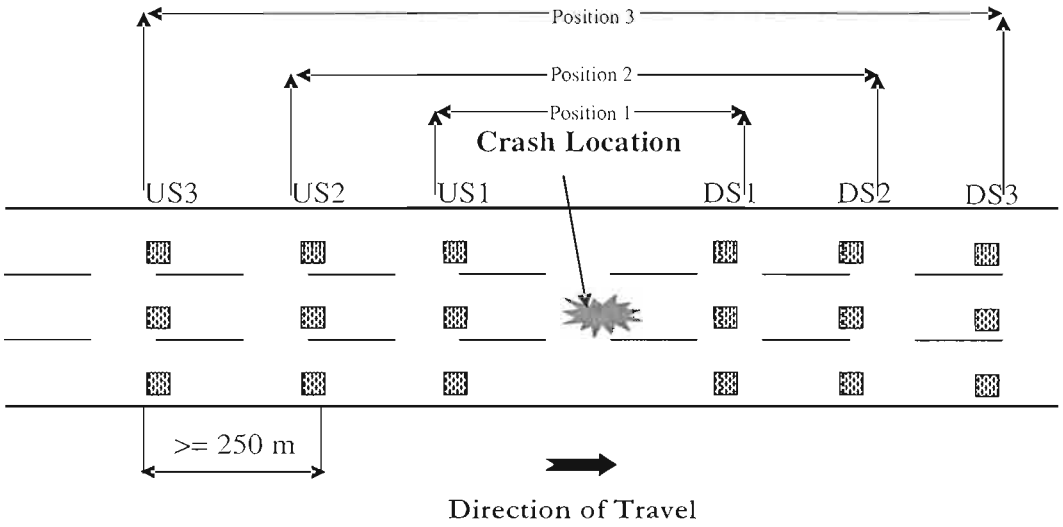

Figure 1: Arrangement of the loop detectors stations

The procedure requires non-crash data to be available corresponding to each crash. For the crash considered in the last paragraph (which was assumed to have occurred at 05:00 p.m. on Monday, September 4, 2006), the corresponding non-crash loop data would be collected for the same time window as the crash data on all Mondays in the month of September 2006. Also, note that these data would be collected from the same upstream and downstream loop detectors stations from which the data corresponding to the crash case were extracted. This sampling scheme controls for other critical factors affecting crash occurrence such as driver population, location on the freeway (geometric features etc.), time of the day and day of the week. The variable 'y' will be 0 for all non crash data.

Next step is the loop data aggregation. The raw 1-minute data were noticed to have random noise. Therefore the raw data were aggregated to 5-minute level in order to obtain averages and standard deviations. The 15-minute period for which data were collected was divided into 3 time slices; numbered 0 through 2. The interval between time of crash and 5 minutes after the crash was named as time slice 0; interval between time of crash and 5 minutes prior to it was named time slice 1; interval between 5 and 10 minutes prior to the crash as time slice 2. The traffic parameters were further aggregated across lanes and the averages (and standard deviations) for speed and volumes at 5-minute level were calculated along with the logarithm of coefficient of variation (standard deviation/average). The aggregation across all lanes is necessary as there were instances in which a particular lane's loop detector was not reporting data. Aggregating over all the lanes lead to a lesser number of missing observations in the dataset.

The nomenclature for these average and standard deviations is of the form 'XYZα_β'. 'X' takes the value A or S for average and standard deviation, respectively; while 'Y' takes the value S or

V for speed and volume, respectively. 'Zα' takes the value of U1, U2, U3 or D1, D2, D3 depending on the station to which a traffic parameter belongs (nearest upstream/downstream station relative to the crash location being U1/D1 and subsequent detectors being U2/D2 and U3/D3, respectively). 'β' takes up the values 0, 1, or 2 referring to aforementioned three time slices. Hence, 'ASD1_2' and 'AVU1_2' represent *average speed* on *station DS1* over *time slice 2* and *average volume* on station *US1* over *time slice 2*, respectively. The corresponding names for coefficient of variation variables can be deduced by dropping the first alphabet (A or S) and replacing it with the term 'CV'.

As mentioned earlier, in addition to speed and volume information traffic surveillance data from these freeways also include a congestion indicator for the location of each loop detector. The flag for congestion is contingent on the speed data being observed with average speed below 50 kmph indicating congestion. Also note that the speed data used to flag locations for congestion are smoothened using exponential smoothing algorithm. Smoothing algorithm ensures that individual observations with relatively high or low values (sometimes referred to as spikes) do not cause the congestion indicator to fluctuate unrealistically. Theoretical details and advantages of the exponential smoothing algorithm may be found in Hunter (1986). A variable 'CON' was created and valued 1 for congested condition and 0 for non-congested conditions in order to convert the flag into a numerical variable. The average and standard deviations for this variable were then created using the same procedure. The nomenclature for average and standard deviations of 'CON' was also identical. For example, the variable 'ACOND1_2' represented average congestion at the loop detector station 'D1' during the (5-minute) time slice 2. Dutch freeways also have a variable speed limit (VSL) system in place. The VSL information disseminated on the DMS coinciding with the pavement location consisting of the loop detector is also part of the traffic surveillance database. For the research problem at hand it is important to find out the following; first, whether variable speed limit was in effect at a particular station/time slice and second, if there had been a change in the values of the variable speed limit applied at those locations within a particular time slice, i.e., a change within a change. Hence, two binary variables, 'V' and 'CW', were created for each of the six stations and three time slices. For example 'V_D1_2'=1 implies that variable speed limit has been implemented on the location of 'DS1' during time slice 2 and 'CW_U1_1'=0 implies that there is no change in the implementation strategy on the location of 'US1' during time slice 1 and so on. Thus there will be one row of data for each crash/non-crash case after the aggregation process with 180 (10 parameters*6 stations*3 time slices) potential input variables. The final dataset had 288 crash and 968 non-crash cases.

The Random Forest based variable selection procedure used in this study requires the datasets to have all corresponding variables to be non-missing. As expected, it was observed that there were a significant number of observations for which at least one of the six stations (for which the data were being collected) was not reporting data. It resulted in some variables with missing values for those observations. Hence, there was a risk of reducing the number of good observations drastically if we use data from all six stations in the same model. We had encountered a similar problem in one of our previous studies (Abdel-Aty et al., 2005). To illustrate this problem caused by failure(s) in loop detectors; let us assume that the probability of failure for each of the six stations is *'a'* and that their failure(s) are independent of each other. The expected proportion of complete cases (i.e., cases with no missing data from any stations) will be $(1\text{-}a)^k$; where *'k'* is the number of stations parameters from which are included in the same model. For $a = 0.15$ (15%

probability of failure) and k=6 there would only be 38% complete observations. But if only 1 station is considered (k=1), then on average 85% of the observations will be complete. Note that this illustration does not provide actual estimate of good observations but just exemplifies the problem of missing observations caused by using data from more stations.

Based on the discussion above it was decided that data from no more than two stations would be used in the same model. Observing traffic parameters from one station at a time could have led to more complete records. However, with data from only one loop detector station there would be no way to examine the interplay between traffic data being observed upstream and downstream of the crash location and the effect it has on the crash risk. Keeping these constraints in perspective the following strategy was adopted for the analysis: The variable selection procedure would be carried out in three independent phases. These three phases differ from each other in the sets of two loop detectors which provide the data used in each phase. The first phase used data corresponding to non-missing observations from set of loop detector referred to as Position 1 in Figure 1 (i.e., stations 'US1' and 'DS1'). The second phase includes data from set of two stations referred to as Position 2 in Figure 1 (i.e., stations 'US2' and 'DS2'). Similarly data corresponding to non-missing observations from the loop detector locations 'US3' and 'DS3' was used in the third phase. The individual phases of this approach would allow us to investigate the impact of traffic patterns observed at loop detector stations located upstream and downstream of the crash location. On the other hand, the difference between the results of the three phases would help in inferring about change in the effect of the traffic parameters on crash risk with respect to time and space. The datasets hence prepared for the three phases had 144, 162, and 143 complete crash records for Positions 1, 2, and 3, respectively. The number of corresponding complete non-crash records was 506, 560, and 505, respectively.

This three phased analysis was conducted for all three time slices. However, the focus of discussion presented in this paper is on variables calculated for time slice 2, i.e. 5-10 minutes prior to crash. This time period is close enough to the crash time and based on our experience should be able to provide insight into crash prone traffic conditions. Time slice 1, i.e. 0-5 minutes prior to crash will be too close to the crash time to work in real time. Also, time slice 0 is the period 0 to 5 minutes after the crash and hence would be interesting only for incident detection, which of course is not the focus of this research. The variables of importance identified by using Random Forests methodology was first proposed by Breiman (2001).

METHODOLOGY

Random Forest method is an example of data mining analysis. Data mining processes are used to find unsuspected relationships in large 'observational' datasets (Hand et al., 2001). These processes typically involve analysis where the objectives of the data analysis have no bearing on the data collection strategy (i.e., no experimental design). Establishing relationship between loop detectors and crash data which are collected independently of each other is an ideal problem for data mining analysis. The authors previously used data mining processes like classification trees for variable selection (Pande and Abdel-Aty, 2006a; Pande and Abdel-Aty, 2006b).

In this study Random Forests which is a collection of multiple tree classifiers are used for variable selection. A decision tree, with all its simplicity, can be very unstable. In other words, small changes in the input variables might result in large changes in the output. In this regard, Random Forests are more robust variable selection tool. Another advantage of using Random

Forests instead of classification trees is that due to an internal test mechanism one need not divide the input dataset into separate training and validation samples. Variable selection by using Random Forest methodology is implemented through 'randomForest' function implementation in R (Liaw and Wiener, 2002).

Random Forests and Variable Importance Scores

As specified earlier a Random Forest is a collection of tree classifiers. A random subset of variables independently sampled from the input variables is used to grow each constituent tree to the full extent. The resulting classification from each tree is then treated as a vote for the corresponding classification. In this study the binary target is the variable 'y', which takes up value 1 for crash records and 0 for non-crash records. For any input vector the forest chooses that class with the maximum number of votes. The process of growing each constituent tree may be divided into the following steps (Breiman and Cutler, 2007):

- For 'N' number of cases in the training set, a bootstrap sample of 'N' is drawn for growing the tree.
- For 'M' input variables, a constant number of 'm' (m<<M) variables are selected at random from 'M' at each node. The best split among the 'm' is used to split the node.
- Trees are grown to the full extent without pruning.

The forest error rate is directly proportional to the correlation between any two trees and inversely proportional to the strength of individual trees. In other words, Random Forests with strong individual trees providing independent information would lead to better classification performance. Random Forests run efficiently on large datasets since they can handle large number of variables without over-fitting the data. Since we are using it as a data exploration tool the feature of interest in this study is its ability to identify variables most significantly associated with the binary target (Breiman and Cutler, 2007).

When a particular tree is grown from a bootstrap sample, one-third of the training cases are left out and not used in the growth of the tree. The left-out cases are called out-of-bag (OOB) data. The OOB cases, effectively an internal test dataset, are used to obtain an unbiased error estimate as well as the estimates of variable importance. The process for assessing importance of variable 'p' in the context of binary classification is as follows:

- First the OOB cases are subjected to every constituent tree grown to the full extent and the votes for the correct class are counted.
- Then the values of the variable 'p' are permuted randomly and the permuted cases are put through the tree again and the votes are again counted.
- The raw importance score of variable 'p' would be the average difference of the votes between the permuted OOB data and the untouched OOB data, across all trees in the forest (Breiman and Cutler, 2007).

Another variable important measure is based on the Gini importance. In this measure whenever a split of a node is made on the variable 'p', the Gini impurity for the two descendant nodes is less than the parent node. The reductions in Gini impurity are added up for each individual variable over all the trees in the forest. It provides an importance score that is generally consistent with permutation importance score. In this study the second variable importance score based on the Gini impurity criterion has been used to assess the importance of variables.

ANALYSIS AND RESULTS

As described earlier in the paper, the variable selection procedure was carried out in three independent phases. These phases differ in the sets of loop detectors from which the traffic data are collected. These sets are referred to as Position 1, Position 2 and Position 3 (see Figure 1) based on the relative position of constituent stations with respect to the crash location. The dataset include averages, standard deviations and the logarithm of the coefficients of variation of speed and volume. The coefficient of variation essentially represents both average and standard deviation. Therefore, it was decided that two separate runs of the variable selection procedure would be carried out. One run will include average and standard deviation as inputs and the other will include coefficient of variation and not the constituent average and standard deviation. This will give 6 sets of important variables (3 positional phases*2 runs).

The datasets were read in R and Random Forests were grown using 'randomForest' function (Liaw and Wiener, 2002). Figures 2(a) and 2(b) below show the variable importance plots for the two runs for the first phase which include data from Position 1 i.e., non-missing observations from stations 'US1' and 'DS1'. The variable importance is based on the mean decrease in Gini. Since this is part of the exploratory analysis the Gini measure is preferable over the actual classification accuracy. The 'Gini' plots provided in Figure 2 demonstrate a clear distinction between the variables which may be important compared to those which might not based on a significant drop in the importance measure.

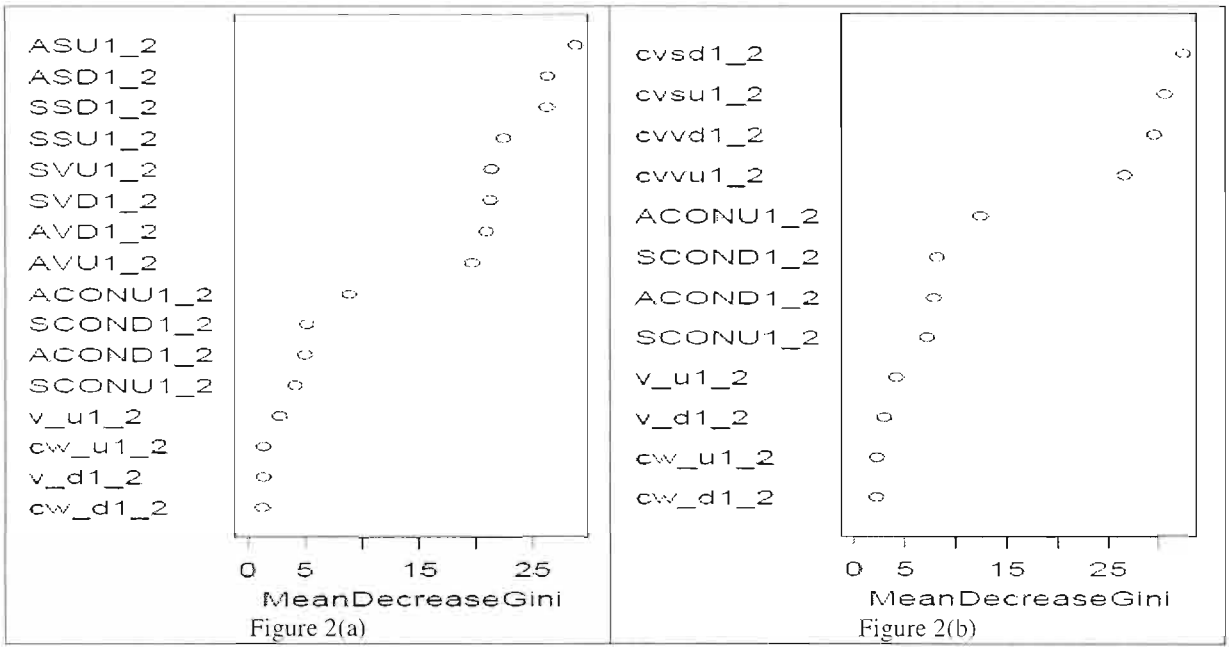

Figure 2: Variable Importance Plots

Figure 2(a) shows the results for the run which includes averages and standard deviations of speeds and volumes along with the variables representing VSL implementation (e.g., V_U1_2). It can be clearly seen that the first eight variables have distinctly higher variable importance score than the remaining variables. Hence, it may be inferred that the actual averages and standard deviations of traffic parameters measured during time slice 2 relate with crash risk more

significantly relative to the variables representing congestion flag and VSL implementation. Figure 2(b) illustrates the results for the run which include logarithm of the coefficients of variation of speed and volume and they are found to be significant for both upstream and downstream loop detector locations. Similar results were found for the other two positional phases (using data from Position 2 and Position 3) for the time slice 2. Table 1 shows the variables of importance (in the order of their importance) selected for all the 3 positional phases for each of the three time slices.

Table 1: Variables of importance for separating crash vs. non-crash cases based on Random Forest

Phase	Time Slice	Important Variables
Position1	Slice 0	ASU1_0, ASD1_0, SSU1_0, AVD1_0, AVU1_0, SSD1_0, SVD1_0, SVU1_0, ACONU1_0
Position1	**Slice 1**	**ASU1_1, ASD1_1, AVD1_1, AVU1_1, SSU1_1, SVU1_1, SSD1_1, SVD1_1**
Position1	**Slice 2**	**ASU1_2, ASD1_2, SSD1_2, SSU1_2, SVU1_2, SVD1_2, AVD1_2, AVU1_2**
Position2	Slice 0	ASU2_0, ASD2_0, SVU2_0, AVD2_0, SVD2_0, AVU2_0, SSU2_0, SSD2_0, ACONU2_0
Position2	**Slice 1**	**ASU2_1, AVD2_1, ASD2_1, ACONU2_1, SSU2_1, SSD2_1, AVU2_1, SVD2_1, SVU2_1**
Position2	**Slice 2**	**ASU2_2, ASD2_2, AVD2_2, SSU2_2, SVU2_2, AVU2_2, SVD2_2, SSD2_2**
Position3	Slice 0	ASD3_0, ASU3_0, SVD3_0, SSU3_0, AVD3_0, AVU3_0, SSD3_0, SVU3_0, ACONU3_0
Position3	**Slice 1**	**ASU3_1, ASD3_1, SSU3_1, AVD3_1, SSD3_1, SVD3_1, SVU3_1, AVU3_1**
Position3	**Slice 2**	**ASU3_2, AVD3_2, SSU3_2, ASD3_2, AVU3_2, SVU3_2, SSD3_2, SVD3_2**

A closer examination of the individual trees constituting the forests revealed that higher variation in speed at both upstream and downstream stations as well as lower average speeds upstream (during time slice 2) increases the likelihood of crashes. Results for time slice 1 (5 minutes interval prior to the crash) were similar to the results in time slice 2 except for the significance of variable representing average congestion at the upstream station (ACONU2_1) for Phase 2.

Note that the rows of Table 1 in 'Bold fonts' depict the variables critical for assessing crash risk (time slices 2 and 3) while the significant variables pertaining to time slice 0 represent variables critical for incident detection. The reason being that time slice 0 represents the 5-minute duration *after* the crash. Note that the measure of congestion at the upstream station is significant at all three phases if one considers time slice 0. Though this study primarily focused on the traffic conditions before the crash, this is an interesting result. The significance of congestion indicator upstream of the crash location (0-5 minute after the crash) is expected since these locations experience congestion after the crash has occurred. It also indicates that that the time of crash information used in this study is reliable.

Another interesting aspect of the variables found significant is that a similar set of variables were found to be significantly associated with rear-end crashes (Pande and Abdel-Aty, 2006 a). Rear end crashes are the single most frequent type of crashes on freeways in the US. In the present study the type of crashes are not known. However, similarity of significant variables does indicate that the rear-end crashes likely dominate the crash data on freeways in the Netherlands as well.

CONCLUSIONS AND FUTURE SCOPE

Loop detectors are essential part of freeway traffic surveillance infrastructure around the world. These detectors provide traffic parameters including speed, volume, and/or some measure of density at time intervals as small as 20 seconds. The objective of the research was to identify real-

time traffic parameters (collected from loop detectors) significantly associated with crashes on the freeways in Netherlands. The set of significant variables helps in assessing whether the approach used for real-time crash risk assessment on Interstate–4 can be transferred to these freeways equipped with more extensive traffic surveillance/management systems. One month of crash data from five freeway sections in the Netherlands along with corresponding loop detector data were used in this study. The presented work is also one of the first applications of Random Forest based variable selection procedure in the area of traffic safety. Random Forests essentially combine multiple classification trees in order to construct a more robust classifier.

One of the most important differences in the data used in this study (compared to most freeways in FL) was the information on variable speed limits. The results in this study show that the variables representing averages and standard deviations of speed/volume are more significantly associated with real-time crash risk compared to the variables representing VSL application (e.g., V_U1_2 and CW_D1_2; See Figure 2). It does not necessarily mean that VSL implementation has no impact on crash risk. It does indicate that to clearly identify the impact of VSL implementation one need to relate it with specific types of crashes (e.g., rear-end). The list of significant variables (averages and standard deviations of speed and volume observed 5-10 minutes before the crash on stations upstream and downstream of the crash site) does suggest that the crash type distribution on these freeways might be similar to Interstate-4 (in FL). In other words, the crashes more prevalent on these freeways might be rear-end and lane change related. It is also possible that the effect of the variable speed limit implementation might be reflected in the speed and volume data being reported by the detectors. The issue, however, needs to be investigated through further analysis by relating specific crash types and traffic conditions.

It is also worth pointing out that ITS countermeasures such as variable speed limits may be used to alleviate crash prone conditions. With infrastructure for implementing VSL already in place on the freeways under consideration; VSL strategies specifically tailored towards reducing the measure of crash risk obtained in this study may be more readily evaluated for these freeways. In a broad sense, these strategies would attempt to reduce the temporal speed variance as well as differential between speeds measured upstream and downstream of crash prone locations.

The next step in the research could be to develop advanced classifiers based on the input variables selected by the Random Forests. Advanced classifiers may be designed to incorporate different sampling scheme that explicitly accounts for the road geometry factors. Explicit accounting of these factors would provide clear indication of their impact on real-time crash risk. Also, in spite of the promise of transferability; there remains a significant scope of improvement. The present study only included data for a single month on five freeway sections. The results need to be validated with a larger sample. Furthermore, if the information on crash type is available then specific analysis by types of crash can be carried out. The information can then be used to assess the risk of particular type of crash on the freeway and more specific warnings may be issued to the drivers through variable message signs. In the Netherlands en-route information delivery system is quite advanced compared to the freeways in FL. Therefore, the implementation of proactive traffic management strategies may be easier to achieve.

ACKNOWLEDGEMENT
The authors would like to thank Mr. Ryan Cunningham for his help in the data preparation process.

REFERENCES:

Abdel-Aty, M., Uddin, N., and Pande, A., Split models for predicting multi-vehicle crashes during high speed and low speed operating conditions on freeways. *Transportation Research Record 1908*, 2005, pp. 51-58.

Breiman, L., Random Forests. *Machine Learning*, Vol. 45, No. 1, 2001, pp. 5-32.

Breiman L. and Cutler A., Random Forests, http://www.stat.berkeley.edu/~breiman/RandomForests/ , Accessed March 18, 2007.

Golob T. and Recker W., A method for relating type of crash to traffic flow characteristics on urban freeways. *Transportation Research Part A*, Vol. 38, No. 1, 2004, pp. 53-80.

Hand D., Mannila H., and Smyth P., Principles of Data Mining. *M.I.T Press*, Cambridge, MA, 2001.

Hunter J., The Exponentially Weighted Moving Average, *Journal of Quality Technology*, 18, 1986, pp. 203 -210.

Liaw A. and Wiener M., Classification and regression by randomForest. *R News*, Vol. 2, 2002, pp. 18-22.

Pande A. and Abdel-Aty M., A Comprehensive Analysis of the Relationship between Real-time Traffic Surveillance data and Rear-end Crashes on Freeways. *Transportation Research Record 1953*, 2006 a, pp. 31-40.

Pande A. and Abdel-Aty M., Assessment of freeway traffic parameters leading to lane-change related collisions. *Accident Analysis and Prevention*, Vol. 38, No. 5, 2006 b, pp. 936-948.

Pande A. and Abdel-Aty M., A Multi-Model Framework for Real-Time Crash Risk Assessment, Forthcoming in the *Journal of the Transportation Research Board* (Accepted March 2007).

Pande A., Abdel-Aty M., and Hsia, L., Spatio-temporal variation of risk preceding crash occurrence on freeways. *Transportation Research Record* 1908, 2004, pp. 26-36.